## Data Understanding

The dataset contains customer reviews for hotels along with their review scores which range between 1 and 5 in the Likert scale where higher scores indicate greater satisfaction with the hotel experience. Each row contains the review score given by customers along with textual feedback expressing their opinions and experiences with the hotel. The text contains a mixture of multiple languages including English, Italian, Chinese and others.

## Data Cleaning & Pre-processing

We first load and read the data first. We notice that there are no missing values in the dataset. Since the reviews are provided in multiple languages, we extract only the English reviews for the purpose of this analysis. Then we select a random sample of 2,000 reviews by setting the seed value as 467. This is done so that the results can be reproduced at any future time.

In order to clean the data, the column names were renamed for our better understanding. To analyse the review texts, we first create a corpus for the review column. A sample of the corpus text is shown below in Figure\_21.

A black screen with white text

Description automatically generated

Figure\_: Sample corpus before preprocessing

The content of the corpus was then tokenized into individual words which makes the data more manageable and helps in for extended analysis in future.

The content is then thoroughly pre-processed by removing URLs, joined words, special characters, punctuations, numbers, stop words and whitespaces using the tm\_map() function from the 'tm' package. All content was also converted to lowercase and accented characters were replaced with ASCII encoding to ensure uniformity across the dataset. Finally, the content was lemmatized to reduce words to their root form and also ensures that the lemmatized words make sense. Once all these steps are done, we get the pre-processed text as shown in Figure\_22.

A black background with white text

Description automatically generated

Figure\_: Sample corpus after preprocessing

Finally, the pre-processed text is combined with the corresponding ratings so as to ensure that the data is standardized and ready for sentiment analysis and topic modelling.

## Data Exploration

We now delve into the exploration and visualization of the cleaned data to gain insights into the distribution of sentiment, frequent words, and their associations within the dataset.

We split the dataset into two parts – one is the positive dataset where ratings equal to or greater than 4 has been considered, and the other is the negative dataset which contains remaining data with ratings below 4. Figure\_23 and Figure\_24 below shows the distribution of ratings by positive and negative sentiment of customers.

A graph with numbers and a bar

Description automatically generated

Figure\_: Distribution of ratings

A pie chart with numbers and a number of percentages

Description automatically generated

Figure\_: Distribution of sentiment

Using both positive and negative datasets, we created separate corpus based on the review text column. Then we created document term matrices with term frequency weighting in order to check which terms occurs often. We clean the matrices then by dropping off terms those occur less than one percent in the documents. We then identified the most frequent words in each sentiment category thus providing insights into the common themes or topics discussed in the reviews. A look at Figure\_25 shows that the terms ‘room’, ‘hotel’, ‘good’, and ‘stay’ occur most frequently for both sentiment categories. This likely meant that most customers who provided either a good or bad rating mentioned about these terms in their review.

A close-up of a graph

Description automatically generated

Figure\_: Top 5 frequent terms in DTM

Furthermore, we also tried to analyse associations as to which terms frequently co-occur with a specific term like ‘travel’. This analysis helps uncover the contextual usage of terms and their relationships within the reviews. For instance, when it comes to travel, terms like ‘child’, ‘fan’, ‘complaint’ was highly associated with the negative dataset which had lower ratings, whereas terms like ‘tube’, ‘london’, ‘station’ had more occurrences within the positive rated dataset as seen in Figure\_26.

A screenshot of a computer

Description automatically generated

Figure\_: Negative and positive associations

The word clouds shown in Figure\_27 and Figure\_28 depicts the frequency of words for each sentiment category using both the term frequency (TF) and term frequency-inverse document frequency (TF-IDF) weighting schemes.

*A couple of words on a white background

Description automatically generated*

Figure\_: TF & TF-IDF wordcloud for positive reviews

*A close-up of words

Description automatically generated*

Figure\_: TF & TF-IDF wordcloud for negative reviews

Since TF-IDF considers not only the frequency of a word within the document but also its importance in the context of the entire dataset, hence It downweighs the importance of common words and upweights terms that are unique to a particular document. Thus TF-IDF provides us more context by highlighting words that are more unique to this particular document. We can see terms like ‘great’, ‘location’, ‘hotel’ come up frequently in positive reviews whereas frequent terms in negative reviews include ‘bed’, ‘clean’, ‘small’, etc.

## Sentiment Analysis

In order to gain insights into the sentiments expressed by customers, we conducted sentiment analysis on the complete dataset.

We used three different methods, namely Syuzhet, Bing, and AFINN, to generate sentiment scores for the entire dataset, which is expressed as positive (1), negative (-1), or neutral (0) as shown in Figure\_29.

A screenshot of a computer

Description automatically generated

Figure\_: Syuzhet, Bing, and AFINN sentiment scores

We used the NRC sentiment lexicon to analyse the sentiment expressed in the hotel reviews. This lexicon contains a comprehensive list of words associated with various sentiment categories such as joy, sadness, anger, anticipation, etc. By leveraging this lexicon, we were able to extract the sentiment expressed in each review across multiple dimensions as shown in Figure\_30.

A screenshot of a computer

Description automatically generated

Figure\_: NRC sentiment for reveiews

We also analysed the count and percentage of words associated with each sentiment category through a bar plot in Figure\_31. We see that overall, the sentiments of the customers in the reviews are positive as most words are associated with sentiment categories like positive, trust, and joy.

A graph of a number of bars

Description automatically generated

Figure\_: Sentiment distribution

## Topic Modelling

Topic modelling is a statistical technique used to uncover latent thematic structures within a collection of documents. It allows us to identify key topics or themes that frequently occur in the text data without requiring prior knowledge of the topics.

In our case, topic modelling will help us gain insights into the main themes and issues expressed in hotel reviews. By analysing the dataset, we can identify common topics that customers frequently discuss in their reviews. This will enable us to understand the factors that contribute to positive and negative customer experiences.

In conducting topic modelling for both the positive and negative reviews separately, we follow a systematic process to uncover the underlying themes and topics within the textual data. The steps involved in this process are as follows:

First, we utilize various metrics such as Griffiths2004, CaoJuan2009, Arun2010, and Deveaud2014 to determine the optimal number of topics for the Latent Dirichlet Allocation (LDA) model. In our case, the optimal number of topics for both positive and negative datasets were identified as 7 as we can see from Figure\_32 and Figure\_33 below.

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

Figure\_: Number of topics for positive reviews

A graph of a number of topics

Description automatically generated

Figure\_: Number of topics for negative reviews

Once the optimal number of topics is determined, we fit the LDA model to the dataset and extract phi (probability distributions over terms for each topic), and theta (probability distributions over topics for each document).

We then analyse the LDA matrix to identify the top terms associated with each topic. Based on these top terms, we label each topic accordingly to facilitate interpretation on what the topics relate to. The labels identified from the positive and negative LDA models are shown in Table 1.

|  |  |  |
| --- | --- | --- |
|  | Positive Reviews | Negative Reviews |
| Topic-1 | Hospitality | Guest Communication |
| Topic-2 | Facilities & Services | Reservation & Check-in |
| Topic-3 | Reservation & Check-in | Hotel Staff & Service |
| Topic-4 | Room Facilities | Room Facilities |
| Topic-5 | Location & Accessibility | Room Comfort |
| Topic-6 | Overall Experience | Overall Experience |
| Topic-7 | Staff Friendliness | Location & Accessibility |

Table : Topic Labelling for Positive & Negative reviews

Once the topics are identified, we calculate the topic probabilities for each review. For instance, we can see from Figure\_34 that for the positive reviews dataset, the second review has the highest probability for associating with the topic ‘Hospitality’.

A screenshot of a computer

Description automatically generated

Figure\_: Topic Probabilities for positive reviews

We explore all the topics now by using the interactive visualization tool LDAvis (Sievert and Shirley, n.d.). Figure\_35 and Figure\_36 shows a graphical representation of topics identified from the positive & negative reviews dataset.

A screenshot of a computer

Description automatically generated

Figure\_: LDAvis representation for positive reviews

A screenshot of a computer

Description automatically generated

Figure\_: LDAvis representation for negative reviews

**Identify Top Factors:** Finally, we calculate the average topic proportions across all reviews to identify the top factors contributing to the reviews.

The top 3 factors that affect the satisfaction of the customers as seen in Figure\_37 are:

* Staff Friendliness
* Location & Accessibility
* Reservation & Check-in

A graph of a number of topics

Description automatically generated with medium confidence

Figure\_: Top 3 topics for customer satisfaction

Similarly, the top 3 factors that affect the dissatisfaction of the customers as seen in Figure\_38 are:

* Location & Accessibility
* Room Facilities
* Overall Experience

A graph with red and grey bars

Description automatically generated

Figure\_: Top 3 topics for customer dissatisfaction

This analysis helps in understanding the most prevalent topics and themes expressed in both positive and negative reviews, providing valuable insights for further analysis and decision-making.